UDACITY

NANODEGREE PROGRAM SYLLABUS

Data Engineering

Overview

Learn to design data models, build data warehouses and data lakes, automate data pipelines, and work with massive datasets. At the end of the program, you'll combine your new skills by completing a capstone project.

Students should have intermediate SQL and Python programming skills.

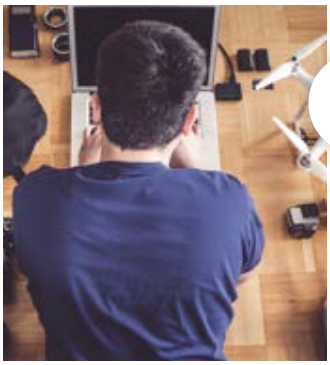
Educational Objectives: Students will learn to

- Create user-friendly relational and NoSQL data models
- Create scalable and efficient data warehouses
- Work efficiently with massive datasets
- Build and interact with a cloud-based data lake
- Automate and monitor data pipelines
- Develop proficiency in Spark, Airflow, and AWS tools

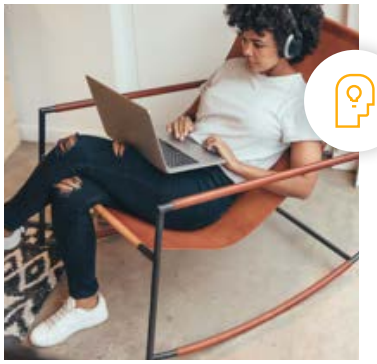
IN COLLABORATION WITH

INSIGHT

Estimated Time:
5 Months at
5 hrs/week



Prerequisites:
Intermediate
Python & SQL



Flexible Learning:
Self-paced, so
you can learn on
the schedule that
works best for you

Need Help?
udacity.com/advisor
Discuss this program
with an enrollment
advisor.

Course 1: Data Modeling

In this course, you'll learn to create relational and NoSQL data models to fit the diverse needs of data consumers. You'll understand the differences between different data models, and how to choose the appropriate data model for a given situation. You'll also build fluency in PostgreSQL and Apache Cassandra.

Course Project
Data Modeling with Postgres

In this project, you'll model user activity data for a music streaming app called Sparkify. You'll create a relational database and ETL pipeline designed to optimize queries for understanding what songs users are listening to. In PostgreSQL you will also define Fact and Dimension tables and insert data into your new tables.

Course Project
Data Modeling with Apache Cassandra

In these projects, you'll model user activity data for a music streaming app called Sparkify. You'll create a database and ETL pipeline, in both Postgres and Apache Cassandra, designed to optimize queries for understanding what songs users are listening to. For PostgreSQL, you will also define Fact and Dimension tables and insert data into your new tables. For Apache Cassandra, you will model your data so you can run specific queries provided by the analytics team at Sparkify.

	LEARNING OUTCOMES	
LESSON ONE	**Introduction to Data Modeling**	• Understand the purpose of data modeling • Identify the strengths and weaknesses of different types of databases and data storage techniques • Create a table in Postgres and Apache Cassandra
LESSON TWO	**Relational Data Models**	• Understand when to use a relational database • Understand the difference between OLAP and OLTP databases • Create normalized data tables • Implement denormalized schemas (e.g. STAR, Snowflake)

LESSON THREE	**NoSQL Data Models**	• Understand when to use NoSQL databases and how they differ from relational databases • Select the appropriate primary key and clustering columns for a given use case • Create a NoSQL database in Apache Cassandran

Course 2: Cloud Data Warehouses

In this course, you'll learn to create cloud-based data warehouses. You'll sharpen your data warehousing skills, deepen your understanding of data infrastructure, and be introduced to data engineering on the cloud using Amazon Web Services (AWS).

Course Project
Build a Cloud Data Warehouse

In this project, you are tasked with building an ELT pipeline that extracts their data from S3, stages them in Redshift, and transforms data into a set of dimensional tables for their analytics team to continue finding insights in what songs their users are listening to.

	LEARNING OUTCOMES	
LESSON ONE	**Introduction to the Data Warehouses**	• Understand Data Warehousing architecture • Run an ETL process to denormalize a database (3NF to Star) • Create an OLAP cube from facts and dimensions • Compare columnar vs. row oriented approaches
LESSON TWO	**Introduction to the Cloud with AWS**	• Understand cloud computing • Create an AWS account and understand their services • Set up Amazon S3, IAM, VPC, EC2, RDS PostgreSQ
LESSON THREE	**Implementing Data Warehouses on AWS**	• Identify components of the Redshift architecture • Run ETL process to extract data from S3 into Redshift • Set up AWS infrastructure using Infrastructure as Code (IaC) • Design an optimized table by selecting the appropriate distribution style and sorting key

Course 3: Spark and Data Lakes

In this course, you will learn more about the big data ecosystem and how to use Spark to work with massive datasets. You'll also learn about how to store big data in a data lake and query it with Spark.

Course Project
Build a Data Lake

In this project, you'll build an ETL pipeline for a data lake. The data resides in S3, in a directory of JSON logs on user activity on the app, as well as a directory with JSON metadata on the songs in the app. You will load data from S3, process the data into analytics tables using Spark, and load them back into S3. You'll deploy this Spark process on a cluster using AWS.

		LEARNING OUTCOMES
LESSON ONE	**The Power of Spark**	• Understand the big data ecosystem • Understand when to use Spark and when not to use it
LESSON TWO	**Data Wrangling with Spark**	• Manipulate data with SparkSQL and Spark Dataframes • Use Spark for ETL purposes
LESSON THREE	**Debugging and Optimization**	• Troubleshoot common errors and optimize their code using the Spark WebUI
LESSON FOUR	**Introduction to Data Lakes**	• Understand the purpose and evolution of data lakes • Implement data lakes on Amazon S3, EMR, Athena, and Amazon Glue • Use Spark to run ELT processes and analytics on data of diverse sources, structures, and vintages • Understand the components and issues of data lakes

Course 4: Automate Data Pipelines

In this course, you'll learn to schedule, automate, and monitor data pipelines using Apache Airflow. You'll learn to run data quality checks, track data lineage, and work with data pipelines in production.

Course Project
Data Pipelines with Airflow

In this project, you'll continue your work on the music streaming company's data infrastructure by creating and automating a set of data pipelines. You'll configure and schedule data pipelines with Airflow and monitor and debug production pipelines.

		LEARNING OUTCOMES
LESSON ONE	**Data Pipelines**	• Create data pipelines with Apache Airflow • Set up task dependencies • Create data connections using hooks
LESSON TWO	**Data Quality**	• Track data lineage • Set up data pipeline schedules • Partition data to optimize pipelines • Write tests to ensure data quality • Backfill data
LESSON THREE	**Production Data Pipelines**	• Build reusable and maintainable pipelines • Build your own Apache Airflow plugins • Implement subDAGs • Set up task boundaries • Monitor data pipelines

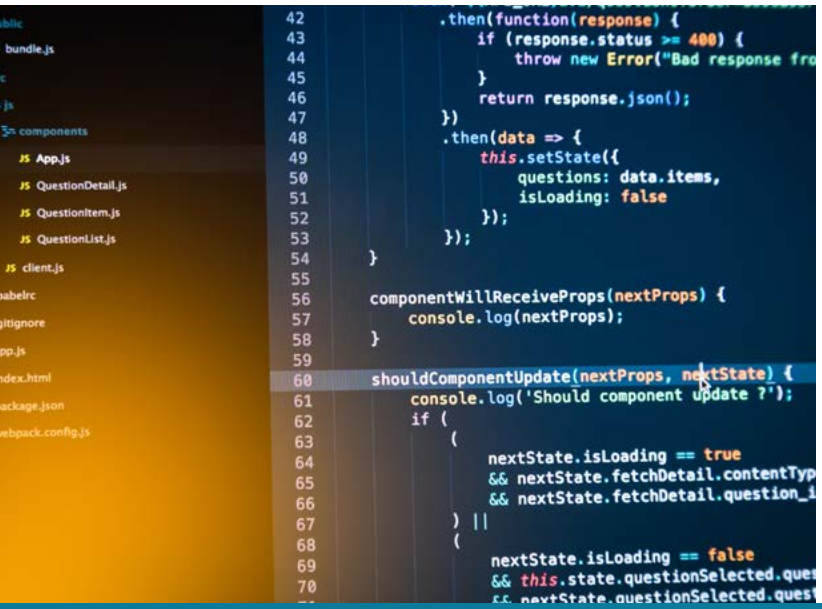
Course 4: Capstone Project

Combine what you've learned throughout the program to build your own data engineering portfolio project.

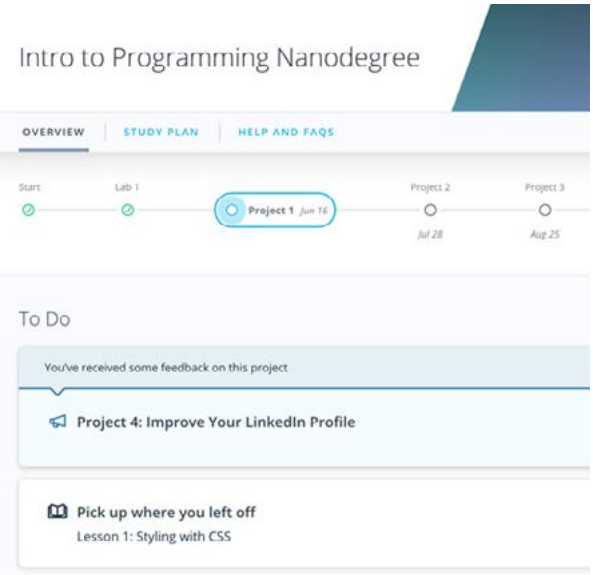
Course Project
Data Engineering Capstone

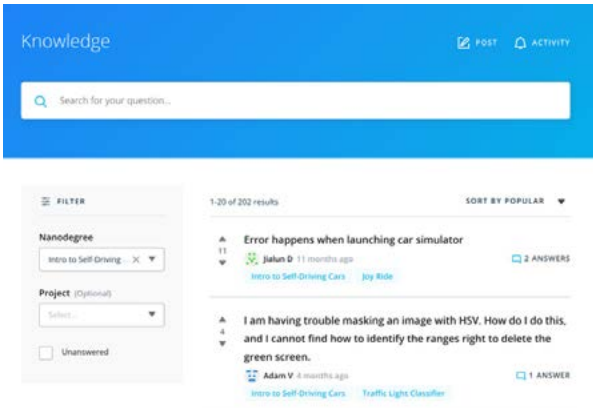
The purpose of the data engineering capstone project is to give you a chance to combine what you've learned throughout the program. This project will be an important part of your portfolio that will help you achieve your data engineering-related career goals.

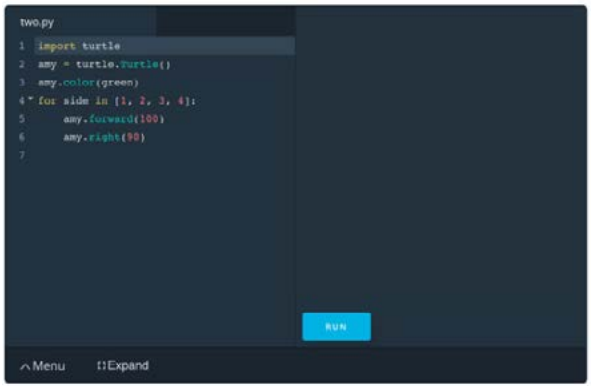
In this project, you'll define the scope of the project and the data you'll be working with. We'll provide guidelines, suggestions, tips, and resources to help you be successful, but your project will be unique to you. You'll gather data from several different data sources; transform, combine, and summarize it; and create a clean database for others to analyze.



Our Classroom Experience







REAL-WORLD PROJECTS

Build your skills through industry-relevant projects. Get personalized feedback from our network of 900+ project reviewers. Our simple interface makes it easy to submit your projects as often as you need and receive unlimited feedback on your work.

KNOWLEDGE

Find answers to your questions with Knowledge, our proprietary wiki. Search questions asked by other students, connect with technical mentors, and discover in real-time how to solve the challenges that you encounter.

WORKSPACES

See your code in action. Check the output and quality of your code by running them on workspaces that are a part of our classroom.

QUIZZES

Check your understanding of concepts learned in the program by answering simple and auto-graded quizzes. Easily go back to the lessons to brush up on concepts anytime you get an answer wrong.

CUSTOM STUDY PLANS

Create a custom study plan to suit your personal needs and use this plan to keep track of your progress toward your goal.

PROGRESS TRACKER

Stay on track to complete your Nanodegree program with useful milestone reminders.

Learn with the Best

Amanda Moran

DEVELOPER ADVOCATE
AT DATASTAX

Amanda is a developer Advocate for DataStax after spending the last 6 years as a Software Engineer on 4 different distributed databases. Her passion is bridging the gap between customers and engineering. She has degrees from University of Washington and Santa Clara University.

Ben Goldberg

STAFF ENGINEER
AT SPOTHERO

In his career as an engineer, Ben Goldberg has worked in fields ranging from Computer Vision to Natural Language Processing. At SpotHero, he founded and built out their Data Engineering team, using Airflow as one of the key technologies.

Luis Serrano

CEO AT NOVELARI & ASSISTANT PROFES-
SOR AT NILE UNIVERSITY

Sameh is the CEO of Novelari, lecturer at Nile University, and the American University in Cairo (AUC) where he lectured on security, distributed systems, software engineering, blockchain and BigData Engineering.

Andrew Paster

DATA ENGINEER
AT WOLT

Olli works as a Data Engineer at Wolt. He has several years of experience on building and managing data pipelines on various data warehousing environments and has been a fan and active user of Apache Airflow since its first incarnations.

Learn with the Best

David Drummond

VP OF ENGINEERING
AT INSIGHT

David is VP of Engineering at Insight where he enjoys breaking down difficult concepts and helping others learn data engineering. David has a PhD in Physics from UC Riverside.

Judit Lantos

DATA ENGINEER
AT SPLIT

Judit was formerly an instructor at Insight Data Science helping software engineers and academic coders transition to DE roles. Currently, she is a Data Engineer at Split where she works on the statistical engine of their full-stack experimentation platform.

Juno Lee

CURRICULUM LEAD
AT UDACITY

Juno is the curriculum lead for the School of Data Science. She has been sharing her passion for data and teaching, building several courses at Udacity. As a data scientist, she built recommendation engines, computer vision and NLP models, and tools to analyze user behavior.

All Our Nanodegree Programs Include:

EXPERIENCED PROJECT REVIEWERS

REVIEWER SERVICES

- Personalized feedback & line by line code reviews
- 1600+ Reviewers with a 4.85/5 average rating
- 3 hour average project review turnaround time
- Unlimited submissions and feedback loops
- Practical tips and industry best practices
- Additional suggested resources to improve

TECHNICAL MENTOR SUPPORT

MENTORSHIP SERVICES

- Questions answered quickly by our team of technical mentors
- 1000+ Mentors with a 4.7/5 average rating
- Support for all your technical questions

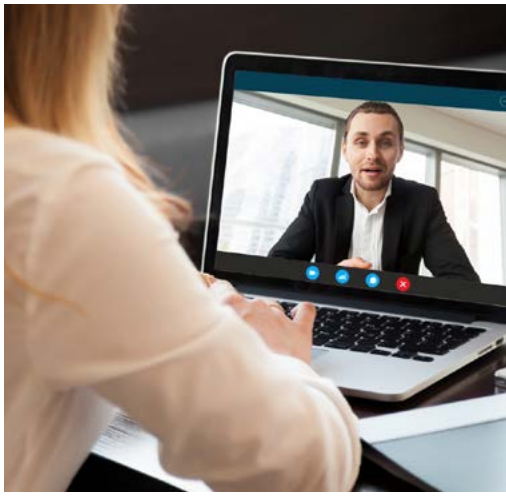
PERSONAL CAREER SERVICES

CAREER SUPPORT

- Resume support
- Github portfolio review
- LinkedIn profile optimization

Frequently Asked Questions

PROGRAM OVERVIEW



WHY SHOULD I ENROLL?

The data engineering field is expected to continue growing rapidly over the next several years, and there's huge demand for data engineers across industries.

Udacity has collaborated with industry professionals to offer a world-class learning experience so you can advance your data engineering career. You will get hands-on experience running data pipelines, building relational and noSQL data models, creating databases on the cloud, and more. Udacity provides high-quality support as you master in-demand skills that will qualify you for high-value jobs in the data engineering field and help you land a job you love.

By the end of the Nanodegree program, you will have an impressive portfolio of real-world projects and valuable hands-on experience.

WHAT JOBS WILL THIS PROGRAM PREPARE ME FOR?

This program is designed to prepare people to become data engineers. This includes job titles such as analytics engineer, big data engineer, data platform engineer, and others. Data engineering skills are also helpful for adjacent roles, such as data analysts, data scientists, machine learning engineers, or software engineers.

HOW DO I KNOW IF THIS PROGRAM IS RIGHT FOR ME?

This Nanodegree program offers an ideal path for experienced programmers to advance their data engineering career. If you enjoy solving important technical challenges and want to learn to work with massive datasets, this is a great way to get hands-on practice with a variety of data engineering principles and techniques.

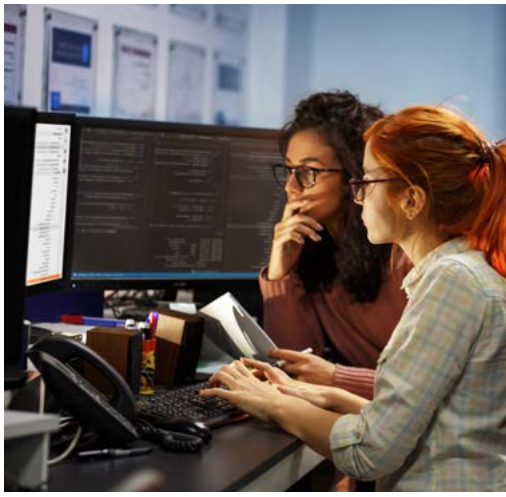
The prerequisites for this program include proficiency in Python and SQL. You should be comfortable writing functions and loops, using classes, working with libraries in Python. You should be comfortable querying data using joins, aggregations, and subqueries in SQL.

WHAT IS THE DIFFERENCE BETWEEN THE DATA ANALYST, MACHINE LEARNING ENGINEER, AND THE DATA SCIENTIST NANODEGREE PROGRAMS?

Udacity's School of Data Science consists of several different Nanodegree programs, each of which offers the opportunity to build data skills, and advance your career. These programs are organized around four main career roles: Business Analyst, Data Analyst, Data Scientist, and Data Engineer.

FAQs Continued

The School of Data currently offers two clearly-defined career paths. These paths are differentiated by whether they focus on developing programming skills or not. Whether you are just getting started in data, are looking to augment your existing skill set with in-demand data skills, or intend to pursue advanced studies and career roles, Udacity's School of Data has the right path for you! Visit "How to Choose the Data Science Program That's Right for You" to learn more.



ENROLLMENT AND ADMISSION

DO I NEED TO APPLY? WHAT ARE THE ADMISSION CRITERIA?

There is no application. This Nanodegree program accepts everyone, regardless of experience and specific background.

WHAT ARE THE PREREQUISITES FOR ENROLLMENT?

The Data Engineer Nanodegree program is designed for students with intermediate Python and SQL skills.

In order to successfully complete the program, students should be comfortable with the following programing concepts:

- Strings, numbers, and variables
- Statements, operators, and expressions
- Lists, tuples, and dictionaries
- Conditions, loops
- Procedures, objects, modules, and libraries
- Troubleshooting and debugging
- Research & documentation
- Problem-solving
- Algorithms and data structures
- Joins
- Aggregations
- Subqueries
- Table definition and manipulation (Create, Update, Insert, Alter)

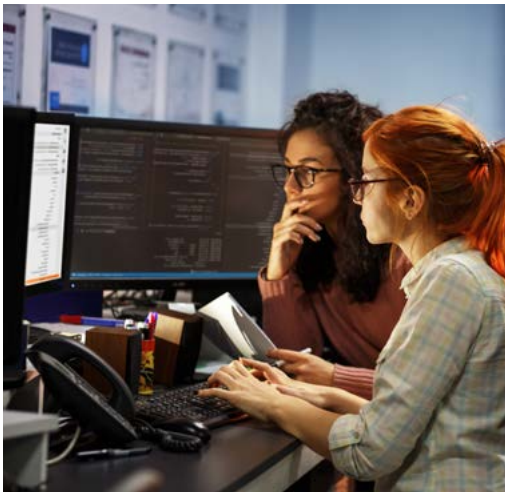
IF I DO NOT MEET THE REQUIREMENTS TO ENROLL, WHAT SHOULD I DO?

Udacity's **Programming for Data Science Nanodegree** program is great preparation for the Data Engineer Nanodegree program. You'll learn to code with Python and SQL.

You can also prepare by taking a number of Udacity's free courses, such as: **Introduction to Python Programming SQL for Data Analysis.**

FAQs Continued

TUITION AND TERM OF PROGRAM



HOW IS THIS NANODEGREE PROGRAM STRUCTURED?

The Data Engineer Nanodegree program is comprised of content and curriculum to support six (6) projects. We estimate that students can complete the program in five (5) months working 10 hours per week.

Each project will be reviewed by the Udacity reviewer network. Feedback will be provided and if you do not pass the project, you will be asked to resubmit the project until it passes.

HOW LONG IS THIS NANODEGREE PROGRAM?

Access to this Nanodegree program runs for the length of time specified in the payment card above. If you do not graduate within that time period, you will continue learning with month to month payments. See the **Terms of Use** and **FAQs** for other policies regarding the terms of access to our Nanodegree programs.

SOFTWARE AND HARDWARE

WHAT SOFTWARE AND VERSIONS WILL I NEED IN THIS PROGRAM?

There are no software and version requirements to complete this Nanodegree program. All coursework and projects can be done via Student Workspaces in the Udacity online classroom.